Known Unknowns and Unknown Unknowns: Designing a Scalable Adverse Event Reporting System for AI

Lindsey A. Gailmard, Drew Spence, Christie Lawrence, Daniel E. Ho

Stanford University

gailmard@stanford.edu, dkspence@stanford.edu, christiem.lawrence@gmail.com, dho@law.stanford.edu

Abstract

There continues to be substantial uncertainty surrounding the risks posed by advanced general-purpose or 'frontier' AI models. Many risks—like security vulnerabilities, misuse, or ethical failures—are highly context-dependent and may only become apparent post-deployment, limiting the feasibility of developing effective ex ante safeguards like pre-deployment testing and capability evaluations. We argue that adverse event (AE) reporting systems, long-used in sectors like healthcare and transportation, offer a scalable and pragmatic solution to this governance gap. AE reporting systems enable continuous monitoring by collecting structured incident data from developers and downstream users, surfacing emergent risks, and supporting adaptive policy responses—providing a path to move from voluntary and ad hoc ethics frameworks to enforceable regulation. To be effective, however, an AE reporting system for AI must align stakeholders' incentives, scale efficiently, and integrate with government infrastructure. This paper helps bridge the divide between the high-level policy priorities and ground-level developers and users as AE reporting systems start to be built out. We motivate our discussion by identifying key challenges for AI regulation that AE reporting seeks to address.

1 Introduction

Artificial intelligence (AI) systems are increasingly deployed in high-stakes domains such as healthcare (Shaheen 2021), criminal justice (Sushina and Sobenin 2020), finance (Cao 2022; Challoumis 2024), and transportation (Bharadiya 2023). The rapid advancement of general-purpose AI technology has significantly expanded its applicability within those domains. In healthcare, for instance, AI tools are being used to support clinical decision-making, including disease diagnosis (Kurtzweil 2024) and treatment, as well as to streamline administrative tasks like documentation and transcription (Bongurala et al. 2024). Similarly, lawyers have used legal AI tools to assist with document preparation (Vayadande et al. 2024) and, in some jurisdictions, courts have even begun exploring the use of AI models to automate administrative tasks and draft judicial opinions (Corren 2024; Mendizabal 2024).

Despite these promising applications, the deployment of AI systems has raised significant concerns about their reliability, safety, and broader social impact. Recent lawsuits have alleged a variety of harms associated with the deployment of AI systems, such as erroneous denials of insurance claims for critical patient care (Mole 2023), damaging and dangerous chatbot dialogue (*e.g.*, recommending violence) (Allyn 2024), and price-fixing in rental markets (Picciotto 2025). Meanwhile, legal AI tools designed to assist with legal research hallucinate at high rates, despite claims that the tools are "hallucination-free" (Magesh et al. 2024).

While these harms are well documented, there continues to be considerable uncertainty regarding the full scope of risks posed by AI systems. Much of this uncertainty stems from the rapid evolution of AI capabilities—particularly among so-called 'frontier models'[1]—and the wide range of contexts AI is deployed. Experts, policymakers, and industry leaders continue to raise alarms about the potential harms of advanced AI technology and frontier models—pointing to potentially catastrophic consequences of broad deployment stemming from malicious and improper use or poor performance, like autonomous weapons development or bioterrorism (Hendrycks, Mazeika, and Woodside 2023).

While there is impetus for AI to be regulated (see, e.g., Meltzer 2023; UN News 2024; McClain et al. 2025), there is significant disagreement about *how* to do so (Economist 2023). Proposals range from FDA-style 'approval' regulation (Carpenter 2014; Carpenter and Ezell 2024) to forms of licensing, disclosure, or pre-deployment audits. Critiques of licensing, registration, disclosure, and auditing emphasize a persistent misalignment between regulatory objectives (e.g., the harms intended to address) and the efficacy of such regulatory proposals in achieving those objectives (Guha et al. 2023). Instead of focusing on the form of regulation, industry stakeholders have endorsed 'risk-based' approaches (Google 2024; Anthropic 2024), though operationalization of those approaches remains ambiguous. Debates about the most appropriate regulation and efforts to develop actionable proposals are constrained by a fundamental obstacle: the absence of a standardized mechanism for tracking and evaluating the real-world use and consequences of AI deployment. This 'data vacuum' (Narayanan and Kapoor 2023) makes it difficult to distinguish speculative concerns from

[1]For a discussion of distinct regulatory challenges posed by frontier models see Anderljung et al. (2023).

substantiated harms (Nordström 2022; Vermeer 2025) and to design regulatory interventions that effectively mitigate harms without unintentionally restricting beneficial uses of AI (Guha et al. 2023).

Policymakers thus face two central questions: (1) how to allocate resources to efficiently and effectively address emergent risks and mitigate future risks associated with AI, and (2) how to avoid stifling innovation, economic competitiveness, or the beneficial adoption of AI technologies. Addressing these questions requires a regulatory framework grounded in empirical evidence and responsive to evolving patterns of use (Bommasani et al. 2025). A potential solution is the creation of post-deployment adverse event (AE) reporting systems—an approach explored in detail by recent scholarship (Guha et al. 2023; Dai et al. 2025). AE reporting systems—long used in public health—serve to identify, analyze, and mitigate harms arising from deployed technologies, including both actual adverse events and near misses (Flink et al. 2005; Connell 2011; NASA 2001).

Forms of AE reporting for AI have gained traction within AI policy discussions (Guha et al. 2023; Dai et al. 2025; National Artificial Intelligence Advisory Committee 2023b) as practical mechanisms to surface risks and advance evidence-based policy. Academic proposals have outlined a range of models for AI-related reporting—from mandated adverse event disclosures (Guha et al. 2023) to broader transparency reporting frameworks (Bommasani et al. 2024) and structured incident reporting mechanisms (Paeth et al. 2024; Agarwal and Nene 2024; Croxton et al. 2024; Frase 2023; Dixon and Frase 2024; Dai et al. 2025) to domain-specific obligations (Kale et al. 2024). Recent legislative proposals have even incorporated incident reporting requirements as part of emerging governance frameworks (Blumenthal and Hawley 2024), but Congress has yet to pass a bill with AE reporting and has not passed any major technology regulation in recent years (Branch and Beller 2025) even on issues that have been met with bipartisan support (Paul 2025).

The clearest policy progress on incident reporting came this July in the White House's AI Action Plan, which directed the federal government to promote AI incident response capacity in the public and private sectors, including through recommended actions like establishing standards and including AI considerations in cybersecurity incident and vulnerability response efforts (White House 2025). Additionally, the AI Action Plan recommended the establishment an AI Information Sharing and Analysis Center (AI-ISAC) to bolster cybersecurity across critical infrastructure sectors by sharing AI-security threat information (White House 2025). Although a positive first step, these efforts do not fully embrace AE reporting, suggesting a need for concrete recommendations on developing an AE reporting system that serves as a pragmatic tool that enables policymakers to learn about risks without completely overhauling existing institutional frameworks.

This paper argues that a post-deployment AE reporting system is not only a critical first step toward building an empirical basis for targeted, adaptive, and effective AI regulation but also a solution to some of the core challenges that have hamstrung regulation. AE reporting systems can help transition voluntary ethics guidelines and ad hoc governance efforts into enforceable, scalable, evidence-based policy by providing regulators, industry stakeholders, and the public with actionable information about real-world risks. When effectively designed, AE reporting facilitates continuous monitoring, supports adaptive policy interventions, and enables the detection of emerging harms.

However, realizing these benefits requires aligning the reporting incentives of developers, deployers, users, and the broader public. Otherwise, even well-intentioned regulatory mandates may result in limited or superficial compliance. This reinforces a central insight: effective and sustainable AE reporting regimes must be *incentive compatible*. Overly broad or ambiguous reporting requirements risk backfiring, either by discouraging participation or overwhelming reporting systems with low-quality or inconsistent data. Drawing on reporting practices in other high-risk domains, we recommend a tiered approach: mandatory reporting should initially be limited to narrowly defined, high-severity adverse events, while voluntary reporting can capture lower-severity harms to lower compliance costs, build trust, and encourage meaningful disclosures.

The remainder of this paper proceeds as follows. First, we outline key challenges facing U.S. AI regulation. We then define AE reporting and examine applications of AE reporting in other sectors. This background offers instructive insights for designing AI-specific AE reporting systems. Finally, we distill design principles from these systems and evaluate tradeoffs in reporting scope, participation mandates, and data access strategies. We argue that integrating these could help address core challenges facing current regulatory efforts and call on both policymakers and industry stakeholders to explore AE reporting for AI.

2 Challenges for AI Regulation

To frame our discussion, we first identify four core challenges that, we argue, currently hinder the development of effective AI regulation in the United States. First, the rapidly evolving capabilities of general-purpose AI systems introduce significant uncertainty around both present and future risks—-including both *known unknowns*, such as how a model might behave when deployed in high-stakes or sensitive contexts, and *unknown unknowns*, such as the emergence of novel capabilities or unforeseen deployment contexts. Second, regulators face information asymmetries relative to industry about potential risks, limiting regulators' ability to effectively target real risks. For example, while Anthropic CEO Dario Amodei has publicly raised concerns about biosecurity risks posed by advanced AI systems (Amodei 2023), the supporting evidence—derived from internal studies involving domain experts—has not been made publicly available. Third, the political economy of AI regulation often positions safety as in tension with innovation, giving industry incentives to resist regulation or lobby for regulation as a form of protection against potential competition. Fourth, the lack of dedicated resources for implementation at both the federal and state level means resource-intensive policy proposals may be unrealistic, at best, and counterproductive, at worst.

Unknown Risks AI systems, particularly general-purpose models, can behave unpredictably in real-world settings (Schaeffer et al. 2025), complicating pre-deployment risk assessments and impeding the development of effective regulatory frameworks (G'Sell 2024). This unpredictability is compounded by the lack of transparency—including intentional obfuscation to protect asserted trade secrets—in how these models operate, making it difficult for regulators to establish clear guidelines. Downstream actors who fine-tune or adapt general-purpose models for specific applications can inadvertently introduce new risks or amplify existing ones (Kim et al. 2025), such as compromising safety features (Zhan et al. 2024), degrading performance (Betley et al. 2025), or enabling misuse.

Regulating downstream developers, however, is challenging due to the diversity and number of actors involved. In this context, traditional regulatory approaches may thus be impractical or risk stifling innovation through overly burdensome regulation. Alternative strategies, such as imposing obligations on upstream developers to mitigate downstream risks or clarifying the application of existing laws to these new contexts (Dudley 2024) have been proposed to address these issues. Yet, the inherent uncertainty in predicting how AI technologies will evolve and be used complicates regulation (Marchant 2011). To apply existing regulatory authorities to governance of new technology requires, first and foremost, an understanding of the evolving risks.

A central challenge in AI governance is the uncertainty as to whether existing regulatory authorities are adequate to address emerging and evolving risks posed by AI systems. In many cases, existing regulatory frameworks may already provide the necessary tools to address AI-related risks, but regulators require better information about the contexts where those risks arise. For instance, agencies such as the Food and Drug Administration (FDA), the Federal Trade Commission (FTC), and the National Highway Traffic Safety Administration (NHTSA) already possess recall authority and enforcement powers that apply to unsafe or deceptive technologies, including AI-enabled systems. The White House's 2023 Blueprint for an AI Bill of Rights[2] also underscores that many AI risks—such as algorithmic discrimination or product safety failures—can be addressed under current consumer protection, civil rights, and public health laws (White House, OSTP 2023).

Information Asymmetries Policymakers and regulators operate at a significant informational disadvantage relative to AI developers. These asymmetries stem from limited in-house expertise, poor visibility into proprietary systems, and industry use of trade secrecy to shield information that may be critical to designing effective regulation. Existing disclosure regimes—such as those overseen by the Securities and Exchange Commission (SEC), Federal Trade Commission (FTC), and National Highway Traffic Safety Administration (NHTSA)—offer only fragmented and often superficial insight into AI-related risks specific to their relevant sectors. For example, while publicly traded companies may disclose AI-related risks in SEC filings, such disclosures tend to be high-level, strategic, and lack technical specificity (Kuhn 2022; Lu 2020). Additionally, companies may engage in 'AI washing,' overstating or misrepresenting the scope and sophistication of their AI capabilities to investors, consumers, and regulators (Barrios et al. 2024). This distorts the regulatory landscape by obscuring the true nature and distribution of risk. In the absence of consistent reporting mechanisms, policymakers are left with an incomplete and often lagging picture of the AI ecosystem.

Politics of Regulation and Industry Capture Debates over AI regulation reflect broader ideological disagreements about open innovation, free enterprise, and government intrusion. Recent legislative proposals have advocated for a laissez-faire regulatory approach to prioritize innovation and avoid stifling the still nascent industry (Liu 2023). The recently-passed One Big Beautiful Bill originally included a ten-year moratorium on state-passed AI laws (Hendrix 2025; Perrigo and Chow 2025). Other legislation has emphasized ethical considerations, civil rights, and national security, pushing for stricter oversight (Lund et al. 2025). These divergent priorities have, in some instances, delayed legislative progress as legislators seek more information about the technology and its capabilities (Kang and Satariano 2023) or focus on catastrophic risks (Green-Lowe 2024). Moreover, lobbying by powerful tech firms can stall progress, as companies push for self-regulation (Hadfield and Clark 2023) or lobby against enforceable constraints (Kang 2025). The absence of a universally accepted definition of AI further complicates regulatory efforts, as it hampers the ability to delineate the scope of regulations and identify which technologies should be subject to oversight. Moreover, the rapid pace of AI development outruns the slower processes of legislation and regulation, creating a 'pacing problem' where laws and regulations lag behind technological advancements (Marchant 2011). Without concrete information about AI risks, ideological commitments may make it difficult to craft unified, proactive policies that address the full spectrum of AI-related risks while also avoiding overburdening industry.

Regulatory Resources Finally, regulatory frameworks require institutional capacity to be effectively implemented. Many of the federal agencies charged with overseeing technology—like the FTC, the Food and Drug Administration (FDA), and the National Institute of Standards and Technology (NIST)—face significant resource and staffing constraints (Nihill 2024). The Center for AI Standards and Innovation (CAISI) (formerly the U.S. AI Safety Institute), created to serve as a central technical authority for AI oversight, remains underfunded and understaffed (National Artificial Intelligence Advisory Committee 2023a). At the state level, resource limitations are even more pronounced. State legislatures, which have recently taken the lead on AI regulation (DePaula et al. 2024; Norden 2025)—proposing policies ranging from algorithmic transparency requirements to bans on biometric surveillance technologies—largely operate without access to dedicated technical advisors or domain-specific policy staff (Jones, Kaye, and Fineberg 2025).

[2]We note that the current status of recommendations and guidance issued by the previous administration is unclear.

The resource gap between regulators and well-funded tech companies means that enforcement lags far behind innovation (Nordström 2022). The U.S. AI Safety Institute (AISI) had a budget of $10 million for fiscal year 2024 (Wilson 2024). In contrast, Anthropic has secured significant funding, including a $4 billion investment from Amazon (Anthropic 2024). Without adequate resources to support implementation—including personnel, expertise, and technical systems—even well-designed regulatory frameworks may fall short of achieving their policy goals (Zakrzewski 2024; Lawrence, Cui, and Ho 2023; Fox-Sowell 2024). For instance, to comply with E.O. 14110's requirement that agencies appoint a Chief AI Officer many agencies tapped officials with significant existing responsibilities (*e.g.*, Chief Information Officer, Chief Data Officer) (Wang et al. 2025).

Forms of licensing, auditing, and active monitoring systems require upfront investments that may be de-prioritized in austere budget environments (e.g., personnel, technology, infrastructure). This financial pressure exacerbates the existing expertise gap between the public and private sectors, allowing industry to outpace regulation leading to unchecked development and deployment just as the capabilities of frontier models are rapidly increasing.

3 Adverse Event Reporting Regulation

3.1 Description

Adverse event reporting systems, as used in this paper, refer to passive monitoring or surveillance schemes that are administered by or in conjunction with government entities (*e.g.*, agencies) to centrally collect specified information about relevant events or incidents[3] from various mandated or voluntary reporters. We use reporters to refer to individuals, organizations, or other entities that provide the specified information either voluntarily (*i.e.*, voluntary reporters) or as required by law or regulation (*i.e.*, mandated reporters).

Developing an AE system requires specifying:

- **Reportable events**: what events must be reported and/or may be reported;
- **Reportable information and reporting process**: what information is reported about the events and how to report that information;
- **Reporting entities**: which entities are required to report and/or may report an event and/or specified information about that event;
- **Reporting timeline**: when entities are required to report events and/or the specified information;
- **Information sharing**: which entities receive information about the events and how that information is used; and
- **Penalties and protections**: any sanctions or mechanisms used to induce compliance.

By collecting reports of damaging events, AE reporting systems improve regulators' abilities to learn about risks in a systematic way. By requiring private-sector actors (*e.g.*, organizations, industry, companies, manufacturers) to report unanticipated or harmful events, AE reporting systems mitigate at least one challenge to AI regulation—information asymmetries between regulators and industry. Due to their involvement in product development, testing, or use, private-sector actors generally have more information about the risks and harms associated with a particular product than regulators. This information is valuable to regulators that face substantial uncertainty about the risks of new and rapidly evolving technology.

International bodies, including OECD, UNESCO, and the IEEE, have all released AI ethics and governance guidelines emphasizing transparency, accountability, and harm mitigation (UNESCO 2021; Corrêa et al. 2023). AE reporting aligns directly with these principles, offering a concrete method to transform values into practice. Rather than waiting for litigation or catastrophic failure, AE reporting systems promote a culture of learning, encourage responsible development, and improve public awareness of risks.

Forms of AE reporting (Guha et al. 2023; Dai et al. 2025) have been proposed as a structured method for capturing and analyzing context-specific, emergent, and low-frequency but high-severity risks—such as edge-case failures, critical malfunctions, and unintended or unexpected harms that may not be evident during pre-deployment testing. These systems can also surface systematic patterns of misuse and degradation that only become apparent through repeated real-world use.

AE reporting systems have long played a critical role in public safety and regulatory oversight across several high-risk sectors, including in the pharmaceutical, medical device, consumer product, transportation, and digital security sectors. We now discuss AE reporting systems used to promote public health, transportation safety, and digital safety.

3.2 Applications for Public Health

For decades, the federal government has used forms of AE reporting to proactively detect and respond to product, pharmaceutical, or device risks by capturing real-world data from industry practitioners, frontline workers, and the public. Early mandatory AE reporting systems were developed to identify public health risks (Flink et al. 2005) in response to public health crises that revealed significant gaps in safety oversight, particularly in pharmaceuticals and medical devices. The Elixir Sulfanilamide disaster (Wax 1995), in which more than 100 people died after ingesting the drug that had been distributed without any safety testing (Paine 2017), prompted a dramatic shift in the U.S. government's approach to drug regulation. Congress passed the Food, Drug, and Cosmetic Act (FDCA) of 1938, which required manufacturers to provide evidence of drug safety before marketing, laying the groundwork for future post-market pharmaceutical surveillance systems. The concept of adverse event reporting became more formalized with the 1962 Kefauver-Harris Amendments, which mandated that pharmaceutical companies report adverse drug reactions to the FDA and established efficacy requirements (Greene and Podolsky 2012). Though not yet formalized or centralized, the reporting requirements demonstrated the value of post-

[3]We use the term "events" going forward to encompass both events and incidents.

market surveillance in uncovering rare but serious harms that informed broader regulatory paradigms that emphasized continuous learning, responsiveness, and evidence-based intervention.

In the decades that have followed, the FDA continues to use AE reporting schemes to monitor for medical and public health risks (Gliklich RE 2014). For example, the FDA currently administers several mandatory and voluntary reporting systems, including FDA's Adverse Event Reporting System (FAERS), its Medical Device Reporting (MDR) program, and MedWatch to monitor post-market safety of FDA-regulated products and identify emerging risks through structured reporting from both manufacturers and healthcare professionals (FDA 2025). FAERS and MDR enable the FDA to monitor for risks without specifying those risks ex ante. Risks identified through the FAERS system, for instance, range from medical device malfunctions, adverse drug reactions, and user difficulties (*e.g.*, poor instructions or medication labeling). The MedWatch system requires the reporting of serious adverse events related to pharmaceuticals and medical devices. This framework has contributed significantly to identifying unsafe products, issuing recalls, and guiding regulatory decisions.

3.3 Applications for Transportation Safety

In the transportation sector, Congress has authorized agencies to collect certain information and mandated disclosures by certain private-sector entities so that agencies could reduce traffic accidents, deaths, and injuries (Wansley 2023). For example, Congress amended the National Traffic and Motor Vehicle Safety Act in 1974 (Public Law 93-492) to require manufacturers to report defects that endangered motor safety, which supported the National Highway Traffic Safety Administration's (NHTSA) ability to set safety standards for vehicles and equipment (Wansley 2023). In 2000, Congress passed the Transportation Recall Enhancement, Accountability, and Documentation (TREAD) Act (Public Law 106–414) which mandated the creation of NHTSA's Early Warning Reporting program—vehicle manufacturers must report information related to defects, injury, or death related to their products and NHTSA is to use such information to identify defects and inform its safety-related rulemaking (Wansley 2023). Similarly, aviation authorities, such as the FAA, mandate incident reporting from airlines and aircraft manufacturers, helping reduce fatalities and near-misses through continuous learning. NASA, for example, administers the Aviation Safety Reporting System (ASRS) for near-miss and safety incidents in collaboration with the FAA (Connell 2011).

3.4 Applications for Digital Safety

More recently, incident reporting systems have been introduced to monitor for cybersecurity threats–including data breaches (NCSL 2022; Harris and General 2016) and AI incidents (Paeth et al. 2024). In fact, as of 2022, all 50 states had laws requiring private businesses to notify individuals of security breaches involving the release of personally identifiable information (NCSL 2022). While the fundamental problems regulators are attempting to resolve resemble those in other non-AI domains, certain considerations and challenges are specific to AI, including definitional issues (National AI Advisory Committee 2023), rapid technological change, and varied and widespread adoption (Bick, Blandin, and Deming 2024).

Cybersecurity provides particularly relevant lessons, as monitoring systems were built over years of experimentation, often by leveraging pre-existing expertise from non-governmental institutions. Following the 1988 Morris worm—the first large-scale cyberattack to severely disrupt internet-connected systems, including those used by the military and major research universities—national attention turned urgently to the vulnerabilities of emerging digital networks. The worm caused significant damage and highlighted the need for a coordinated approach to cybersecurity threats (Spafford 1989). In response, the Defense Advanced Research Projects Agency (DARPA) collaborated with Carnegie Mellon University to create the Computer Emergency Response Team Coordination Center (CERT/CC) within the Software Engineering Institute (SEI), aiming to research software bugs impacting security and to collaborate with businesses and government entities to enhance overall software and internet security. This innovative partnership combined academic research capabilities with government resources to address a rapidly evolving cyber threat landscape and set a precedent for coordinated incident response that endures today (CERT/CC, 2021).

The Cyber Incident Reporting for Critical Infrastructure Act (CIRCIA) of 2022 includes mandatory reporting requirements (CISA 2024a), whereby covered entities (CISA 2024c) must report relevant cyber incidents within 72 hours of becoming aware of the incident. Similarly, Europe's General Data Protection Regulation (GDPR) introduced mandatory breach reporting for data controllers (of the European Union 2016). The policy requires the disclosure of personal data breaches within 72 hours, ensuring transparency and user protection (Board 2023). Likewise, the EU Artificial Intelligence Act, includes provisions that require high-risk AI systems to have human oversight, continuous monitoring, and post-market data collection. The Act specifically mentions the need to notify authorities about any serious incidents or malfunctions that violate fundamental rights or result in harm.

4 Designing an Adverse Event Reporting System for AI

We now turn to our findings and recommendations about design features that should be considered and included when developing an AE reporting system for AI. We first surveyed existing studies of AE reporting systems in the three safety contexts discussed above—(1) public health, (2) transportation safety, and (3) digital safety—and the effectiveness of those systems to identify our recommended design features. Table 2: Select studies of US reporting systems in the Appendix compiles and summarizes key findings from **21** such studies. We identified five specific design features that policymakers should prioritize in designing an AE reporting system for AI: reportable events, continuous evaluation, penal-

ties and liability protections, mandatory and voluntary reporting, and data access.

After identifying these design features, we surveyed the presence of those design features in AE reporting systems administered by or in collaboration with U.S. federal or state governments that monitor risks in consumer products or that arise from the interactions between human decision processes and technical systems (*i.e.*, our three identified safety contexts). Table 1 lists 11 of these reporting systems and some of our key findings about the design features of those systems. Table 3 (in the Appendix) provides an overview of other AE reporting systems administered by the federal government, including emergency reporting systems and other reporting systems unrelated to product or technology safety. Our analysis of these AE reporting systems demonstrates that well-defined and standardized AE reporting mechanisms not only improve safety and accountability but are feasible and enforceable in complex technological systems. We discuss below how the five design features address the four core challenges for AI regulation that we discussed in section 2.

4.1 Reportable Events

For an AE reporting system to be effective, it must begin with a clear and narrowly defined set of reporting criteria that specifies the types of events to be reported.[4] Narrowly scoped criteria in the early stages help ensure that the reports submitted are both consistent and actionable, minimizing noise and maximizing the system's utility for early risk detection and policy development. As emphasized in the literature on incident reporting systems in high-risk domains, overly broad criteria can dilute the informational quality of submissions and overwhelm agencies with low-priority reports, making it harder to identify genuine safety threats (Leveson 2011; Crumpler and Lewis 2019) and generate actionable insights. Starting with narrowly tailored definitions allows administering agencies and entities to refine operational processes and build capacity before scaling up reporting requirements when necessary.

Initial reporting frameworks for an AE reporting for AI should target high-severity incidents—such as safety-critical failures, emergent malicious behavior, or security breaches—where there is a clear public or systemic risk. This approach is consistent with best practices in other sectors, including the FDA Sentinel Initiative, which began with narrowly focused adverse event criteria before expanding as the infrastructure matured (Platt et al. 2018; U.S. Food and Drug Administration 2008, 2015). Categorizing adverse events can help with identifying which events to initially focus reporting requirements. For example, the FDA categorizes adverse events according to the severity and seriousness of the impact of the event (*e.g.*, whether it results in death or permanent incapacity), whether the event was unexpected (*i.e.*, has not previously been observed or documented), and whether the adverse event is determined (*e.g.*, through implication or affirmative evidence) to be caused by to the drug or medical device (Gliklich, Dreyer, and Leavy 2014; Lucas et al. 2022). Narrowly defined reporting requirements focus analysis and follow-up on serious events (Donaldson, Corrigan, and Kohn 2000). As discussed in section 4.2, reporting systems should incorporate regular reviews, as the AI landscape evolves, to strike a balance between responsiveness and efficiency, ensuring that the system adapts to capture emerging challenges and a wider range of risks, as necessary, without being overloaded.

[4]We discuss the entities responsible for reporting AE, which must also be clear and narrowly defined, in section 4.3.

4.2 Continuous Evaluation

Reporting Requirements An iterative approach to adverse event reporting requirements is essential for keeping pace with evolving technological risks. Static or overly prescriptive criteria risk becoming obsolete as new threats emerge, while adaptive frameworks enable regulators to refine reporting obligations in response to empirical trends and technological change. This approach is already used in other sectors. For example, California's data breach notification law has undergone several amendments (Harris and General 2016) to account for new types of sensitive data, such as biometric identifiers, that were not previously considered high-risk (California Civil Code § 1798.82). These amendments reflect the state's responsiveness to emerging data security threats and provide a model for an iterative approach to AI safety reporting.

This flexibility is crucial in the context of AI, where the risk landscape is still poorly understood and continues to shift rapidly with new model capabilities and deployment practices. Iterative regulatory design—such as sunset clauses that specify an expiration date for mandates, periodic review mandates that may require assessing the adequacy of reporting criteria or the continued relevance of regulations (Bennear and Wiener 2021), or rulemaking authority for agencies—allows legal frameworks to adapt without requiring full statutory overhaul. Scholars of adaptive governance emphasize that iterative mechanisms improve policy responsiveness and reduce the risks of regulatory mismatch or lag (Bennear and Wiener 2019; Craig et al. 2017). Legislators play a key role in this process by ensuring agencies have the legal authority and flexibility to update definitions of reportable events, adjust enforcement mechanisms, and incorporate feedback from both regulated entities and independent researchers.

Agency Authorities Adverse event reporting systems can highlight gaps in regulatory authority by providing regulators and lawmakers with real-world evidence of how AI is used—and misused—in practice. This continuous flow of information could enable agencies to assess whether current legal and regulatory tools are sufficient or whether Congress must provide agencies with new authorities to regulate real risks identified through AE reporting. Evidence-driven evaluation of the sufficiency of current regulatory authorities is essential to avoid both under- and over-regulation. On one hand, AE data can identify policy blind spots by surfacing previously unknown harms and edge-case failures. On the other hand, AE data may demonstrate that certain harms are not as numerous as expected. Thus, AE data could support

System	Context	Mandatory or Voluntary	Liability Protections	Public Data Access	Administering Agency/Entity
FDA Averse Event Reporting System (FAERS)	Drug safety	Mandatory for manufacturers; voluntary for healthcare providers and consumers	Yes for voluntary reports: civil liability shield for providers	Yes	FDA
Vaccine Adverse Event Reporting System (VAERS)	Vaccine safety	Voluntary (some mandates for providers under emergency use)	Yes: civil liability shield for both manufacturers and providers	Yes	FDA; CDC
MedWatch	Drug and device safety	Mandatory for manufacturers; voluntary for consumers and providers	For voluntary reports: civil liability shield for providers	Yes	FDA
Vaccine Safety Datalink (VSD)	Vaccine safety	Voluntary	Limited (aggregate summaries)	Yes	CDC
Aviation Safety Reporting System (ASRS)	Aviation safety	Voluntary	Yes: civil liability shield for reporters and immunity from FAA enforcement actions including flying certificate suspensions	No (anonymized reports available)	FAA; NASA
Near Midair Collision System (NMACS)	Aviation safety	Mandatory for pilots/Air Traffic Controllers	Limited, use de-identified reports for analysis	Limited (aggregate summaries)	FAA
Early Warning Reporting	Autonomous vehicle safety	Mandatory for manufacturers	Reports may be anonymized	Yes	NHTSA
Cyber Incident Reporting (CIRCIA)	Cybersecurity	Voluntary and mandatory (in process)	In process	Limited during investigations	CISA
CERT Coordination Center & Vulnerability Notes Database	Cybersecurity	Voluntary	No	Yes	SEI, Carnegie Mellon University (with federal funding); CISA
California Data Security Breach Reporting	Consumer data protection	Mandatory for entities with breaches affecting Californians	No	Yes	California Department of Justice
AI Incident Database (AIID)	AI/ML systems	Voluntary	No formal protections	Yes	Non-Governmental (Partnership on AI, AI Incident Database project)

Table 1: Examples of Adverse Event Reporting Systems and Description of Design Features

targeted and adaptive legislative and regulatory action (U.S. Department of Transportation, NHTSA 2023).

AE reporting infrastructure would support these agencies by enabling early warnings, pattern recognition, and sector-specific insights needed to operationalize existing mandates effectively (Dudley 2024). In this way, reporting not only supports the case for new authorities where necessary but also enhances the strategic use of existing regulatory capacities.

At the same time, rather than duplicating efforts or creating siloed reporting structures, AI-related incidents should be reported through extensions of existing systems when possible, with cross-referencing capabilities to flag incidents implicating other agencies. This approach can streamline reporting for regulated entities, avoid fragmentation, and enhance coordination between agencies. For example, AI-related failures in medical devices should be routed through the FDA's MedWatch or FAERS systems, while AI incidents in aviation should leverage the ASRS. Coordination across agencies, possibly facilitated through a centralized AI office

or interagency teams, can ensure that AI-specific risks are both detected in their application domains. This integrated model supports regulatory efficiency and helps prevent critical AI harms from falling through jurisdictional gaps.

Proactive Mitigation Adverse event reporting systems are essential components of a life-cycle approach to AI governance—one that links ex ante evaluation (i.e., pre-deployment testing and risk assessments) with ex post monitoring. While ex ante tools aim to identify and mitigate foreseeable risks before deployment, AE reporting captures failures that only become evident in real-world contexts. This feedback loop between pre-deployment testing and post-deployment monitoring enables continuous learning and iterative risk management throughout a system's operational life.

Reporting creates shared visibility into real-world failures or near-misses, allowing private sector developers and downstream deployers to learn from others' experiences and proactively address similar risks in their own systems. By surfacing recurring patterns or critical edge cases, this transparency distributes responsibility for risk mitigation that supports iterative model improvement. This principle of continuous learning is central to modern safety engineering and has been increasingly emphasized in AI ethics literature as key to responsible development (Morley et al. 2020, 2021).

Regulatory precedents reinforce this approach. Evidence from Vaccine Safety Datalink (VSD) demonstrates the value of continuous sequential analysis to enhance the sped and power of signal detection (Huang, Moon, and Segal 2014; Silva and Kulldorff 2015; Kulldorff and Silva 2017). Similarly, NHTSA has implemented a near-real-time reporting mandate for automated driving system (ADS) developers, requiring the submission of accident data within one day of occurrence (Wansley 2023). These disclosures enable adaptive safety governance through the rapid identification of vehicle software vulnerabilities and allow for timely recall notices. In addition, the Cybersecurity and Infrastructure Security Agency (CISA) plays a central role in managing federal cyber incident reporting and vulnerability mitigation efforts, despite its relatively modest budget of under $3 billion (Crumpler and Lewis 2024). Through its Continuous Diagnostics and Mitigation (CDM) program, CISA has identified and helped remediate more than 25 million cybersecurity vulnerabilities across federal networks, improving the security of civilian government systems (CISA 2024b). The agency's Vulnerability Warning Pilot, launched in 2022, extended these efforts to the private sector by proactively notifying over 7,600 organizations about vulnerabilities, resulting in the mitigation of more than 3 million vulnerabilities (CISA 2024b).

A life-cycle model could similarly cultivate an industry-wide culture of safety and accountability, encouraging developers to integrate lessons from real-world deployment into continuous model validation, testing, and retraining efforts. This dynamic, adaptive approach is critical for ensuring that AI systems remain safe and aligned with societal values as they evolve.

Independent Risk Assessments In contrast to regulations that require firms to report self-assessed sources of risk periodically (e.g., SEC quarterly reports), an AE report is triggered by a realized event. This limits provides regulators with access to the underlying risk data and limits discretion.

Adverse event reporting systems would enable regulators to conduct independent, evidence-based risk assessments using granular data about actual harms, malfunctions, or unintended outcomes encountered in real-world settings. These systems have proven effective in other domains—such as aviation, healthcare, and cybersecurity—at uncovering new or systemic risks that were missed in pre-deployment testing or obscured by organizational incentives (Leveson 2016; Sherry et al. 2025; Crumpler and Lewis 2019). Applied to AI, adverse event reports could help surface emergent failure modes, bias amplification, security vulnerabilities, or downstream misuse, all of which are critical to ensuring that regulators and the public are not wholly reliant on industry for risk intelligence.

4.3 Mandatory and Voluntary Reporting

A hybrid reporting approach—combining mandatory and voluntary components—offers a pragmatic and flexible framework for monitoring post-deployment AI risks. This is because, even where AE reporting systems have incentive-compatible reporting mechanisms like liability protections, AE reporting systems that rely on voluntary disclosures may still suffer from under-reporting of events and incomplete information (Robb et al. 2012; U.S. Food and Drug Administration 2008, 2015). A hybrid system could impose mandatory reporting obligations on core AI developers, who possess the most technical insight into system behavior and failure modes, while allowing voluntary reports from downstream users, integrators, and end-users to enrich the overall risk landscape. This structure mirrors successful reporting systems in other sectors. For example, FDA's Medical Device Reporting (MDR) program mandates reports from manufacturers, importers, and device user facilities, while also collecting voluntary submissions from healthcare providers, patients, and consumers to capture a wider range of adverse events and rare occurrences (FDA, 2022).

In the cybersecurity domain, CISA is transitioning to a hybrid model under the forthcoming final rule implementing the Cyber Incident Reporting for Critical Infrastructure Act (CIRCIA) of 2022. CIRCIA requires CISA to establish a national reporting system for significant cyber incidents and ransom payments that impacted sixteen critical infrastructure sectors (*e.g.*, energy, transportation, and financial sectors) (CIRCIA, Public Law 117-103). This system will require timely reporting from covered critical infrastructure entities, as defined in regulation (CISA 2024c), while continuing to accept voluntary reports from any entity that experienced a cyber incident, enhancing situational awareness across both public and private sectors (CISA, 2023). CISA has published fact sheets and other guidance documents detailing who are mandatory reporters, what entities CISA encourages to make voluntary reports, and when incident reporting must be made (CISA 2023). Once implemented, this mandatory reporting framework will enable more targeted

threat mitigation, strengthening the nation's cyber defense capabilities. These efforts underscore the potential of structured reporting systems—not only to reduce vulnerability exposure at scale but also to inform proactive government responses. CISA's model demonstrates how effective risk monitoring and data sharing can significantly improve national security outcomes. Hybrid reporting could encourage faster identification of high-severity risks post-deployment, especially when paired with incentives such as liability protections or safe harbor provisions (Strauss et al. 2025).

4.4 Penalties and Liability Protections

Incentive-compatible reporting mechanisms are essential for the success and integrity of AE reporting systems, especially when applied to AI systems where the risks are complex, emergent, and often opaque. The core challenge is that entities—such as AI developers, deployers, or downstream users—may lack incentives to voluntarily disclose harmful or risky outcomes, particularly if doing so could expose them to legal liability, heightened regulatory scrutiny, reputational damage, or competitive disadvantage. To address this, AE reporting systems must be designed so that participating entities see the value in compliance and do not face disproportionate penalties for honest disclosures.

Mechanisms to ensure incentive compatibility may include safe harbor protections that shield entities from punitive enforcement actions if they promptly and transparently report adverse events, as long as there is no willful negligence or misconduct. Anonymized or confidential reporting channels can encourage participation, especially in early stages of system deployment when the full range of harms is unknown. Furthermore, reciprocal benefits, such as access to aggregated industry data, benchmarking reports, or technical support, can turn reporting into a cooperative rather than adversarial process. For example, some systems, like the FAA's ASRS, allow for anonymous submissions that incentivize the reporting of sensitive or stigmatized information while shielding reporters from punitive consequences—an approach that has significantly improved aviation safety culture (Connell 2011).

4.5 Data Access

Information Sharing Between Agencies Effective adverse event reporting systems should incorporate structured information-sharing protocols that enable relevant agencies with domain-specific expertise to respond quickly and appropriately to reported risks. Centralized intake of reports—with secure distribution to agencies such as the FDA, NHTSA, FTC, or CISA, depending on the sector—can facilitate timely interventions and leverage existing regulatory capabilities. While the protection of proprietary information is essential, especially for reports involving trade secrets or confidential model architectures, mechanisms for anonymized data sharing are critical to incentivize reporting (O'Leary and Chappell 1996). Anonymized summaries, trend analyses, and retrospective review can help prevent repeated failures and foster a shared safety culture across developers and sectors (Flott et al. 2018; Darveau and Hannon 2017).

Precedents from cybersecurity policy offer valuable models for information-sharing structures. Under the CIRCIA reporting system, federal agencies that receive a cyber incident report must share that report with CISA and the Department of Homeland Security must establish an intergovernmental council to coordinate and deconflict incident reporting requirements across the federal government (CISA 2023).

Cross-Sector Collaboration The Joint Cyber Defense Collaborative (JCDC), operated by CISA, brings together public and private sector actors to share real-time threat intelligence and coordinate proactive defense strategies (CISA 2021). Similarly, Information Sharing and Analysis Organizations (ISAOs) enable cross-sector collaboration by disseminating actionable threat data while protecting confidential business information (EO 13691, 2015). Coordinated vulnerability disclosure (CVD) processes, widely used in cybersecurity and critical infrastructure sectors, provide a structured method for engaging developers, vendors, and government stakeholders to jointly fix cyber vulnerabilities before disclosing those vulnerabilities publicly (Householder et al. 2025). AI safety reporting could mirror these practices by establishing anonymized databases, public dashboards, and risk briefings to support shared learning and reduce duplication of harm across the ecosystem (MITRE 2024).

Public Data Access Public access to adverse event reports plays a critical role in supporting both regulatory oversight and independent research aimed at identifying and mitigating uncertain risks—each of the 11 systems surveyed in Table 1 allows for some public data access, even if subject to additional restrictions. The FDA maintains the FDA Adverse Event Reporting System (FAERS), a publicly accessible database that enables researchers to conduct large-scale analyses of medication safety, uncovering potential statistical associations between drugs and adverse events (FDA, 2024). These datasets not only enhance transparency but also empower external experts to validate regulatory findings and identify emerging safety concerns that may not be evident through internal review alone. They also may provide some insight into the frequency of certain types of prompts.

Moreover, open access to adverse event data has spurred innovation in data science and public health. Researchers have applied machine learning and natural language processing techniques to mine FAERS and similar databases for previously undetected patterns, including early warnings of drug interactions or rare adverse reactions (Harpaz et al. 2012). Beyond pharmacovigilance, adverse event data has shown potential utility in broader epidemiological monitoring. For instance, studies suggest that patterns in medication error and adverse event reports can offer early detection of disease outbreaks (Wong et al. 2003) or shifts in healthcare delivery practices (Rafter et al. 2015). These cross-disciplinary applications illustrates the expanding role of transparency in fostering technical advancements to improve health outcomes.

Government-Academic Partnerships The history of cyber incident reporting highlights the strategic value of government-academic partnerships in building resilient national security infrastructure. The federal government's collaboration with CERT/CC has demonstrated how academic institutions can offer vital analytical and operational support, filling technical gaps in government agencies with cutting-edge expertise and agility. Over the years, it has contributed to the discovery and coordination of responses to thousands of software vulnerabilities and played a key role in developing best practices for cyber defense (U.S. Senate Committee on Commerce and Transportation 2018).

In public health, the RADAR (Research on Adverse Drug Events and Reports) project demonstrates the value of partnerships in uncovering previously undetected drug safety issues (Bennett et al. 2007). Established in 1998, the partnership between academic researchers and regulators has successfully identified adverse reactions to cancer therapies and anticoagulants (Bennett et al. 2007) that were not detected during clinical trials or passive surveillance systems. Through independent case investigations and analysis of incident data, RADAR has prompted FDA communications and label changes and influenced clinical practice guidelines (Bennett et al. 2005, 2007). The success of RADAR in surfacing adverse drug reactions more quickly than FDA alone (Bennett et al. 2007) underscores how academic institutions can provide essential analytical capacity to support active surveillance—serving as a scalable template for risk detection in AI. Researchers may also develop novel statistical techniques for analyzing incident data (Bangard et al. 2025; Tan, Markatou, and Chakraborty 2025).

These models of collaboration offer valuable lessons for the design of AI safety oversight. As AI technologies become more pervasive and complex, a similar academic-government framework could support risk analysis, helping federal agencies interpret, prioritize, and respond to emerging risks more efficiently. Leveraging academic expertise in areas such as algorithmic auditing and machine learning could make an AI adverse event reporting system both more scalable and adaptive to technological change.

5 Limitations

We conclude by acknowledging four unresolved challenges that fall outside the scope of this paper, but will be critical for the development of an effective AE reporting system for AI.

First, AE reporting systems are limited in their ability to convey risk magnitude without contextual information about base rates (Singleton et al. 1999), such as model deployment or prompt type frequencies. Without this information, policymakers lack critical information for estimating risk probabilities and assessing causality (Shimabukuro et al. 2015). Access to usage statistics are essential for interpreting the significance of observed incidents and prioritizing regulatory responses accordingly.

Second, the value of AE reports depends on the ability of reporting entities to identify and interpret adverse events. Mandated or voluntary reporting can only yield actionable insights if individuals or organizations are able to identify when a reportable event has occurred—highlighting the importance of clear reporting criteria, easy to use reporting portals, and tiered obligations based on the technical sophistication, organizational scale, or access to model output of reporting entities (Brundage et al. 2024).

Third, any new AE reporting framework for AI must be interoperable with existing systems, such as those used in healthcare, product safety, and cyber incident reporting. A key design consideration is how an AI-specific AE system can be effectively integrated with or complement these existing mechanisms to avoid duplication, reduce reporting burdens, and ensure coordination across regulatory domains.

Finally, while our recommendation to restrict reporting to a narrow set of adverse events, at least initially, could focus scarce regulatory resources on more speculative harms, it aims to balance the informativeness of reports against the burdens of reporting obligations and investigation resources. This does not diminish the importance of existing risk, like bias.

6 Conclusion

Effective AI governance must begin with an understanding of the real-world risks associated with deployed systems. Pre-deployment assessments alone are increasingly insufficient as applications of AI technology grow. AE reporting systems can complement these ex ante mechanisms by capturing failures, unintended consequences, and emergent harms that arise post-deployment—particularly those that are context-specific or difficult to anticipate. These systems enable continuous monitoring and iterative risk assessment, facilitating agile, evidence-based governance.

Policymakers should prioritize the development of AI-specific AE reporting systems to manage AI risks. Much of the necessary infrastructure can be pursued under existing statutory authority and integrated into existing regulatory frameworks. Agencies like the FDA, CDC, NHTSA, and CISA already manage sophisticated post-market surveillance systems that can serve as blueprints for AI AE reporting. Embedding AE reporting within existing institutional practices provides a pragmatic, near-term strategy to mitigate novel risks while broader governance regimes are developed.

The recent action plan for AI released by the White House is a starting point for developing the monitoring capacity and information-sharing infrastructure to assess the safety of AI in real-world settings. However, the proposal should go further by adopting a hybrid reporting model for critical AI events, engaging cross-sector partnerships to support active surveillance, and allowing open access to anonymized incident data to accelerate learning and the development of evidence-based practices for AI risk management.

Industry stakeholders must also support these efforts. Developers and deployers of AI systems bear direct responsibility for the safety and performance of their technologies in society. An AE reporting system can provide critical feedback to improve models and identify failures that evade pre-deployment detection. AI companies can hence also benefit from collaborating around an AI system and the development of shared safety infrastructure.

Acknowledgments

We are grateful to Rishi Bommasani, Sarah Cen, Percy Liang, Caroline Meinhardt, Anka Reuel, Angelina Wang, the AI policy working group at Stanford and the feedback of anonymous reviewers.

References

Agarwal, A.; and Nene, M. J. 2024. Addressing AI Risks in Critical Infrastructure: Formalising the AI Incident Reporting Process. In *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 1–6. IEEE.

Allyn, B. 2024. Lawsuit: A chatbot hinted a kid should kill his parents over screen time limits.

Amodei, D. 2023. Written Testimony of Dario Amodei, Ph.D. Co-Founder and CEO, Anthropic.

Anderljung, M.; Barnhart, J.; Korinek, A.; Leung, J.; O'Keefe, C.; Whittlestone, J.; Avin, S.; Brundage, M.; Bullock, J.; Cass-Beggs, D.; Chang, B.; Collins, T.; Fist, T.; Hadfield, G.; Hayes, A.; Ho, L.; Hooker, S.; Horvitz, E.; Kolt, N.; Schuett, J.; Shavit, Y.; Siddarth, D.; Trager, R.; and Wolf, K. 2023. Frontier AI Regulation: Managing Emerging Risks to Public Safety. arXiv:2307.03718.

Anthropic. 2024. The Case for Targeted Regulation.

Anthropic. 2024. Powering the next generation of AI development with AWS.

Bangard, J.; Holsbø, E.; Svendsen, K.; Perduca, V.; and Birmelé, E. 2025. Detecting adverse high-order drug interactions from individual case safety reports using computational statistics on disproportionality measures. arXiv:2504.00646.

Barrios, J. M.; Campbell, J. L.; Johnson, R. G.; and Liu, Y. C. 2024. Artificially Intelligent or Artificially Inflated? Determinants and Informativeness of Corporate AI Disclosures. *Determinants and Informativeness of Corporate AI Disclosures (August 23, 2024)*.

Bennear, L. S.; and Wiener, J. B. 2019. Adaptive Regulation: Instrument Choice for Policy Learning over Time.

Bennear, L. S.; and Wiener, J. B. 2021. Pursuing Periodic Review of Agency Regulation.

Bennett, C. L.; Nebeker, J. R.; Lyons, E. A.; Samore, M. H.; Feldman, M. D.; McKoy, J. M.; Carson, K. R.; Belknap, S. M.; Trifilio, S. M.; Schumock, G. T.; et al. 2005. The research on adverse drug events and reports (RADAR) project. *Jama*, 293(17): 2131–2140.

Bennett, C. L.; Nebeker, J. R.; Yarnold, P. R.; Tigue, C. C.; Dorr, D. A.; McKoy, J. M.; Edwards, B. J.; Hurdle, J. F.; West, D. P.; Lau, D. T.; et al. 2007. Evaluation of serious adverse drug reactions: a proactive pharmacovigilance program (RADAR) vs safety activities conducted by the Food and Drug Administration and pharmaceutical manufacturers. *Archives of internal medicine*, 167(10): 1041–1049.

Betley, J.; Tan, D.; Warncke, N.; Sztyber-Betley, A.; Bao, X.; Soto, M.; Labenz, N.; and Evans, O. 2025. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*.

Bharadiya, J. P. 2023. Artificial intelligence in transportation systems a critical review. *American Journal of Computing and Engineering*, 6(1): 35–45.

Bick, A.; Blandin, A.; and Deming, D. 2024. The Rapid Adoption of Generative AI.

Blumenthal, R.; and Hawley, J. 2024. Bipartisan Framework for U.S. AI Act.

Board, E. D. P. 2023. Guidelines 9/2022 on personal data breach notification under GDPR.

Bommasani, R.; Arora, S.; Chayes, J.; Choi, Y.; Cuéllar, M.-F.; Fei-Fei, L.; Ho, D. E.; Jurafsky, D.; Koyejo, S.; Lakkaraju, H.; et al. 2025. Advancing science-and evidence-based AI policy. *Science*, 389(6759): 459–461.

Bommasani, R.; Klyman, K.; Longpre, S.; Xiong, B.; Kapoor, S.; Maslej, N.; Narayanan, A.; and Liang, P. 2024. Foundation Model Transparency Reports. arXiv:2402.16268.

Bongurala, A. R.; Save, D.; Virmani, A.; and Kashyap, R. 2024. Transforming health care with artificial intelligence: redefining medical documentation. *Mayo Clinic Proceedings: Digital Health*, 2(3): 342–347.

Branch, J.; and Beller, I. 2025. Buried in Congress's Budget Bill is a Push to Halt AI Oversight.

Brundage, M.; Avin, S.; Clark, J.; Toner, H.; Eckersley, P.; Garfinkel, B.; Dafoe, A.; Scharre, P.; Zeitzoff, T.; Filar, B.; Anderson, H.; Roff, H.; Allen, G. C.; Steinhardt, J.; Flynn, C.; hÉigeartaigh, S. ; Beard, S.; Belfield, H.; Farquhar, S.; Lyle, C.; Crootof, R.; Evans, O.; Page, M.; Bryson, J.; Yampolskiy, R.; and Amodei, D. 2024. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv:1802.07228.

Cao, L. 2022. Ai in finance: challenges, techniques, and opportunities. *ACM Computing Surveys (CSUR)*, 55(3): 1–38.

Carpenter, D. 2014. Reputation and power: Organizational image and pharmaceutical regulation at the FDA. In *Reputation and Power*. Princeton University Press.

Carpenter, D.; and Ezell, C. 2024. An FDA for AI? Pitfalls and Plausibility of Approval Regulation for Frontier Artificial Intelligence. In *AAAI/ACM Conference on AI, Ethics, and Society*, volume 7. San Jose, California: Association for the Advancement of Artificial Intelligence, Association for Computing Machinery.

Challoumis, C. 2024. the landscape of AI in Finance. In *XVII International Scientific Conference*, 109–144.

CISA. 2021. CISA Launches New Joint Cyber Defense Collaborative.

CISA. 2023. Cyber Incident Reporting for Critical Infrastructure Act of 2022 (CIRCIA) Fact Sheet.

CISA. 2024a. Cyber Incident Reporting for Critical Infrastructure Act (CIRCIA) Reporting Requirements.

CISA. 2024b. Opening Statement by CISA Director Jen Easterly.

CISA. 2024c. What Would Be a Covered Entity under CIRCIA as Proposed in 6 CFR § 226.2?

Connell, L. 2011. NASA aviation safety reporting system (ASRS). Technical report.

Corrêa, N. K.; Galvão, C.; Santos, J. W.; Del Pino, C.; Pinto, E. P.; Barbosa, C.; Massmann, D.; Mambrini, R.; Galvão, L.; Terem, E.; et al. 2023. Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, 4(10).

Corren, B. 2024. Chief Justice Creates Task Force on Use of Generative AI in the California Courts.

Craig, R. K.; Garmestani, A. S.; Allen, C. R.; Arnold, C. A. T.; Birgé, H.; DeCaro, D. A.; Fremier, A. K.; Gosnell, H.; and Schlager, E. 2017. Balancing stability and flexibility in adaptive governance: an analysis of tools available in US environmental law. *Ecology and society: a journal of integrative science for resilience and sustainability*, 22(2): 1.

Croxton, J.; Robusto, D.; Thallam, S.; and Calidas, D. 2024. Message Incoming: Establish an AI Incident Reporting System.

Crumpler, W.; and Lewis, J. A. 2019. The Cybersecurity Workforce Gap.

Crumpler, W.; and Lewis, J. A. 2024. Department of Homeland Security.

Dai, J.; Raji, I. D.; Recht, B.; and Chen, I. Y. 2025. Aggregated Individual Reporting for Post-Deployment Evaluation. arXiv:2506.18133.

Darveau, K.; and Hannon, D. 2017. Barriers and facilitators to voluntary reporting and their impact on safety culture. *The International Journal of Aerospace Psychology*, 27(3-4): 92–108.

DePaula, N.; Gao, L.; Mellouli, S.; Luna-Reyes, L. F.; and Harrison, T. M. 2024. Regulating the machine: An exploratory study of US state legislations addressing Artificial Intelligence, 2019-2023. In *Proceedings of the 25th Annual International Conference on Digital Government Research*, 815–826.

Dixon, R. B. L.; and Frase, H. 2024. An Argument for Hybrid AI Incident Reporting: Lessons Learned from Other Incident Reporting Systems. *Georgetown Center for Security and Emerging Technology*.

Donaldson, M. S.; Corrigan, J. M.; and Kohn, L. T. 2000. To err is human: building a safer health system.

Dudley, S. E. 2024. Lessons from the Past for Regulating AI.

Economist, T. 2023. The world wants to regulate AI, but does not quite know how.

FDA. 2025. Public Dashboard: FDA Adverse Event Reporting (FAERS) System.

Flink, E.; Chevalier, L.; Ruperto, A.; Dameron, P.; Heigel, F.; Leslie, R.; Mannion, J.; and Panzer, R. 2005. *Lessons Learned from the Evolution of Mandatory Adverse Event Reporting Systems*.

Flott, K.; Nelson, D.; Moorcroft, T.; Mayer, E. K.; Gage, W.; Redhead, J.; and Darzi, A. W. 2018. Enhancing safety culture through improved incident reporting: a case study in translational research. *Health Affairs*, 37(11): 1797–1804.

Fox-Sowell, S. 2024. Local governments need more cyber funding, report finds.

Frase, H. 2023. AI Incident Collection: An Observational Study of the Great AI Experiment. *Georgetown Center for Security and Emerging Technology*.

Gliklich, R. E.; Dreyer, N. A.; and Leavy, M. B. 2014. Registries for evaluating patient outcomes: a user's guide.

Gliklich RE, L. M., Dreyer NA. 2014. *Registries for Evaluating Patient Outcomes: A User's Guide [Internet]*. Agency for Healthcare Research and Quality (US).

Google. 2024. Recommendations for Regulating AI.

Green-Lowe, J. 2024. CAIP Applauds the Romney-Led, Bipartisan Bill to Address Catastrophic AI Risks.

Greene, J. A.; and Podolsky, S. H. 2012. Reform, regulation, and pharmaceuticals—the Kefauver–Harris Amendments at 50. *New England Journal of Medicine*, 367(16): 1481–1483.

G'Sell, F. 2024. Regulating Under Uncertainty: Governance Options for Generative AI.

Guha, N.; Lawrence, C.; Gailmard, L. A.; Rodolfa, K.; Surani, F.; Bommasani, R.; Raji, I.; Cuéllar, M.-F.; Honigsberg, C.; Liang, P.; and Ho, D. E. 2023. AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing. *George Washington Law Review*.

Hadfield, G. K.; and Clark, J. 2023. Regulatory markets: The future of AI governance. *arXiv preprint arXiv:2304.04914*.

Harpaz, R.; DuMouchel, W.; Shah, N. H.; Madigan, D.; Ryan, P.; and Friedman, C. 2012. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology & Therapeutics*, 91(6): 1010–1021.

Harris, K. D.; and General, A. 2016. California data breach report. *Retrieved August*, 7: 2016.

Hendrix, J. 2025. US Senate Drops Proposed Moratorium on State AI Laws in Budget Vote.

Hendrycks, D.; Mazeika, M.; and Woodside, T. 2023. An Overview of Catastrophic AI Risks. arXiv:2306.12001.

Householder, A. D.; Sarvepalli, V. S.; Havrilla, J.; Churilla, M.; Pons, L.; Lau, S.-h.; VanHoudnos, N. M.; Kompanek, A.; and McIlvenny, L. 2025. https://insights.sei.cmu.edu/blog/protecting-ai-from-the-outside-in-the-case-for-coordinated-vulnerability-disclosure/.

Huang, Y.-L.; Moon, J.; and Segal, J. B. 2014. A comparison of active adverse event surveillance systems worldwide. *Drug safety*, 37(8): 581–596.

Jones, A. C.; Kaye, J. Z.; and Fineberg, H. V. 2025. Supplying State Legislatures With Scientific Expertise.

Kale, A. U.; Dattani, R.; Tabansi, A.; Hogg, H. D. J.; Pearson, R.; Glocker, B.; Golder, S.; Waring, J.; Liu, X.; Moore, D. J.; and Denniston, A. K. 2024. AI as a Medical Device Adverse Event Reporting in Regulatory Databases: Protocol for a Systematic Review. *JMIR Res Protoc*, 13.

Kang, C. 2025. Emboldened by Trump, A.I. Companies Lobby for Fewer Rules.

Kang, C.; and Satariano, A. 2023. As A.I. Booms, Lawmakers Struggle to Understand the Technology.

Kim, M.; Kim, Y.; Kang, H. J.; Seo, H.; Choi, H.; Han, J.; Kee, G.; Park, S.; Ko, S.; Jung, H.; et al. 2025. Fine-Tuning LLMs with Medical Data: Can Safety Be Ensured? *NEJM AI*, 2(1): AIcs2400390.

Kuhn, K. X. 2022. Algorithmic Decision-Making and Corporate Risk: Toward Transparency Through Corporate Disclosures. *Nebraska Law Review*, 100(4): 7.

Kulldorff, M.; and Silva, I. R. 2017. Continuous post-market sequential safety surveillance with minimum events to signal. *Revstat statistical journal*, 15(3): 373.

Kurtzweil, J. 2024. New AI model draws treasure maps to diagnose disease.

Lawrence, C.; Cui, I.; and Ho, D. 2023. The Bureaucratic Challenge to AI Governance: An Empirical Assessment of Implementation at U.S. Federal Agencies. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 606–652. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Leveson, N. G. 2011. The use of safety cases in certification and regulation.

Leveson, N. G. 2016. *Engineering a safer world: Systems thinking applied to safety*. The MIT Press.

Liu, I. 2023. Can AI companies still move fast and break things despite pending regulations.

Lu, S. 2020. Algorithmic opacity, private accountability, and corporate social disclosure in the age of artificial intelligence. *Vand. J. Ent. & Tech. L.*, 23: 99.

Lucas, S.; Ailani, J.; Smith, T. R.; Abdrabboh, A.; Xue, F.; and Navetta, M. S. 2022. Pharmacovigilance: reporting requirements throughout a product's lifecycle. *Therapeutic advances in drug safety*, 13: 1–16.

Lund, B.; Orhan, Z.; Mannuru, N. R.; Bevara, R. V. K.; Porter, B.; Vinaih, M. K.; and Bhaskara, P. 2025. Standards, frameworks, and legislation for artificial intelligence (AI) transparency. *AI and Ethics*, 1–17.

Magesh, V.; Surani, F.; Dahl, M.; Suzgun, M.; Manning, C. D.; and Ho, D. E. 2024. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. arXiv:2405.20362.

Marchant, G. E. 2011. Addressing the pacing problem. *The growing gap between emerging technologies and legal-ethical oversight: the pacing problem*, 199–205.

McClain, C.; Kennedy, B.; Gottfried, J.; Anderson, M.; and Pasquini, G. 2025. How the U.S. Public and AI Experts View Artificial Intelligence.

Meltzer, J. P. 2023. The US government should regulate AI if it wants to lead on international AI governance.

Mendizabal, V. 2024. Courts in Buenos Aires are using ChatGPT to draft rulings.

MITRE. 2024. MITRE Launches AI Incident Sharing Initiative.

Mole, B. 2023. UnitedHealth uses AI model with 90% error rate to deny care, lawsuit alleges.

Morley, J.; Elhalal, A.; Garcia, F.; Kinsey, L.; Mökander, J.; and Floridi, L. 2021. Ethics as a service: a pragmatic operationalisation of AI ethics. *Minds and Machines*, 31(2): 239–256.

Morley, J.; Floridi, L.; Kinsey, L.; and Elhalal, A. 2020. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, 26(4): 2141–2168.

Narayanan, A.; and Kapoor, S. 2023. Generative AI companies must publish transparency reports.

NASA. 2001. ASRS: The Case for Confidential Incident Reporting Systems. Technical report.

National AI Advisory Committee. 2023. Rationales, Mechanisms, and Challenges to Regulating AI: A Concise Guide and Explanation.

National Artificial Intelligence Advisory Committee. 2023a. RECOMMENDATION: Implementation of the NIST AI Safety Institute.

National Artificial Intelligence Advisory Committee. 2023b. RECOMMENDATION: Improve Monitoring of Emerging Risks from AI through Adverse Event Reporting.

NCSL. 2022. Security Breach Notification Laws.

Nihill, C. 2024. FTC modernization, enforcement efforts jeopardized by cuts, officials say.

Norden, L. 2025. States Take the Lead on Regulating Artificial Intelligence.

Nordström, M. 2022. AI under great uncertainty: implications and decision strategies for public policy. *AI & society*, 37(4): 1703–1714.

of the European Union, O. J. 2016. General Data Protection Regulation (GDPR).

O'Leary, M.; and Chappell, S. L. 1996. Confidential incident reporting systems create vital awareness of safety problems. *ICAO journal*, 51(8): 11–13.

Paeth, K.; Atherton, D.; Pittaras, N.; Frase, H.; and McGregor, S. 2024. Lessons for Editors of AI Incidents from the AI Incident Database. arXiv:2409.16425.

Paine, M. F. 2017. Therapeutic disasters that hastened safety testing of new drugs.

Paul, K. 2025. Parents are desperate to protect kids on social media. Why did the US let a safety bill die?

Perrigo, B.; and Chow, A. R. 2025. Senators Reject 10-Year Ban on State-Level AI Regulation, In Blow to Big Tech.

Picciotto, R. 2025. U.S. Expands RealPage Price-Fixing Lawsuit to Include Six Big Landlords.

Platt, R.; Brown, J. S.; Robb, M.; McClellan, M.; Ball, R.; Nguyen, M. D.; and Sherman, R. E. 2018. The FDA Sentinel Initiative—an evolving national resource. *N Engl J Med*, 379(22): 2091–2093.

Rafter, N.; Hickey, A.; Condell, S.; Conroy, R.; O'connor, P.; Vaughan, D.; and Williams, D. 2015. Adverse events in healthcare: learning from mistakes. *QJM: An International Journal of Medicine*, 108(4): 273–277.

Robb, M. A.; Racoosin, J. A.; Sherman, R. E.; Gross, T. P.; Ball, R.; Reichman, M. E.; Midthun, K.; and Woodcock, J. 2012. The US Food and Drug Administration's Sentinel Initiative: expanding the horizons of medical product safety. *Pharmacoepidemiology and drug safety*, 21(1): 9.

Schaeffer, R.; Schoelkopf, H.; Miranda, B.; Mukobi, G.; Madan, V.; Ibrahim, A.; Bradley, H.; Biderman, S.; and Koyejo, S. 2025. Why Has Predicting Downstream Capabilities of Frontier AI Models with Scale Remained Elusive? arXiv:2406.04391.

Shaheen, M. Y. 2021. Applications of Artificial Intelligence (AI) in healthcare: A review. *ScienceOpen Preprints*.

Sherry, L.; Raz, A.; Amissah, M.; and Shortle, J. 2025. Integrating System Safety Analysis (SSA) AND SysML: Opportunities and Challenges.

Shimabukuro, T. T.; Nguyen, M.; Martin, D.; and DeStefano, F. 2015. Safety monitoring in the vaccine adverse event reporting system (VAERS). *Vaccine*, 33(36): 4398–4405.

Silva, I. R.; and Kulldorff, M. 2015. Continuous versus group sequential analysis for post-market drug and vaccine safety surveillance. *Biometrics*, 71(3): 851–858.

Singleton, J. A.; Lloyd, J. C.; Mootrey, G. T.; Salive, M. E.; Chen, R. T.; Ellenberg, S.; Rastogi, S.; Krueger, C.; Braun, M.; Wise, R.; et al. 1999. An overview of the vaccine adverse event reporting system (VAERS) as a surveillance system. *Vaccine*, 17(22): 2908–2917.

Spafford, E. H. 1989. The internet worm incident. In *European Software Engineering Conference*, 446–468. Springer.

Strauss, I.; Moure, I.; O'Reilly, T.; and Rosenblat, S. 2025. Real-World Gaps in AI Governance Research. *arXiv preprint arXiv:2505.00174*.

Sushina, T.; and Sobenin, A. 2020. Artificial intelligence in the criminal justice system: leading trends and possibilities. In *6th International Conference on Social, economic, and academic leadership (ICSEAL-6-2019)*, 432–437. Atlantis Press.

Tan, Y.; Markatou, M.; and Chakraborty, S. 2025. Flexible Empirical Bayesian Approaches to Pharmacovigilance for Simultaneous Signal Detection and Signal Strength Estimation in Spontaneous Reporting Systems Data. arXiv:2502.09816.

UN News. 2024. 'Irrefutable' need for global regulation of AI: UN experts.

UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence.

U.S. Department of Transportation, NHTSA. 2023. Report to Congress: Proposed Improvements to Early Warning Reporting Data.

U.S. Food and Drug Administration. 2008. The Sentinel Initiative National Strategy for Monitoring Medical Product Safety. Technical report.

U.S. Food and Drug Administration. 2015. Sentinel Program Interim Assessment (FY 15). Technical report.

U.S. Senate Committee on Commerce, S.; and Transportation. 2018. Written Testimony of Art Manion, Hearing on "Complex Cybersecurity Vulnerabilities: Lessons Learned from Spectre and Meltdown".

Vayadande, K.; Bhat, A.; Bachhav, P.; Bhoyar, A.; Charoliya, Z.; and Chavan, A. 2024. AI-Powered Legal Documentation Assistant. In *2024 4th International Conference on Pervasive Computing and Social Networking (ICPCSN)*, 84–91. IEEE.

Vermeer, M. J. D. 2025. Could AI Really Kill Off Humans?

Wang, J. W.; Suzgun, M.; Meinhardt, C.; Zhang, D.; Nowacki, K.; and Ho, D. E. 2025. Assessing the Implementation of Federal AI Leadership and Compliance Mandates.

Wansley, M. T. 2023. Regulating driving automation safety. *Emory LJ*, 73: 505.

Wax, P. M. 1995. Elixirs, diluents, and the passage of the 1938 Federal Food, Drug and Cosmetic Act. *Annals of internal medicine*, 122(6): 456–461.

White House. 2025. Winning the Race: America's AI Action Plan.

White House, OSTP. 2023. Blueprint for an AI Bill of Rights.

Wilson, C. 2024. The US Has Committed to Spend Far Less Than Peers on AI Safety.

Wong, W.-K.; Moore, A. W.; Cooper, G. F.; and Wagner, M. M. 2003. Bayesian network anomaly pattern detection for disease outbreaks. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, 808–815.

Zakrzewski, C. 2024. This agency is tasked with keeping AI safe. Its offices are crumbling.

Zhan, Q.; Fang, R.; Bindu, R.; Gupta, A.; Hashimoto, T.; and Kang, D. 2024. Removing RLHF Protections in GPT-4 via Fine-Tuning. arXiv:2311.05553.